The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata

Konstantinos Liolios[1], Konstantinos Mavromatis[2], Nektarios Tavernarakis[3], and Nikos C. Kyrpides[2]

[1]University of Chicago, Chicago, USA, [2]Genome Biology Program, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, USA, [3]Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology, Heraklion, Crete, Greece.

Corresponding Author:

Nikos C. Kyrpides

DOE Joint Genome Institute

2800 Mitchell Drive, B400

Walnut Creek, CA 94598

Phone: 925-296-5718

Fax: 925-296-5850

Email: NCKyrpides@lbl.gov

ABSTRACT

The Genomes On Line Database (GOLD) is a comprehensive resource of information for genome and metagenome projects world-wide. GOLD provides access to complete and ongoing projects and their associated metadata through pre-computed lists and a search page. The database currently incorporates information for more than 2900 sequencing projects, of which 639 have been completed and the data deposited in the public databases. GOLD is constantly expanding to provide metadata information related to the project and the organism and is compliant with the Minimum Information about a Genome Sequence" (MIGS) specifications. GOLD is available at http://www.genomesonline.org and also mirrored at the Institute of Molecular Biology and Biotechnology, Crete, Greece at http://gold.imbb.forth.gr/

HISTORY AND GROWTH

Since its instigation in 1997, GOLD (1, 2, 3) has been constantly monitoring genome sequencing projects worldwide and providing the community with a unique centralized database integrating diverse information related to Archaeal, Bacterial, Eukaryotic and more recently Metagenomic sequencing projects.

In contrast to what was anticipated in the previous report of the database two years ago (3), the total number of identified projects has not yet doubled, currently reaching 2905 (compared to 1575 on September 2005). However, if only the archaeal and bacterial projects would be considered, then the total current number is reaching 1950 projects, only 36 projects short from doubling the number in two years. The advent of new sequencing technologies, such as pyrosequencing (4), has certainly significantly contributed to the continuous increase in the rate of new microbial sequencing projects. In fact 134 GOLD projects are now reported using 454 technology as part of the Whole Genome Sequencing (WGS) project.

Two major large scale microbial genome sequencing programs have been launched during the last two years which also account for the majority of the reported 454 sequencing projects. The first is the Human Gut Microbiome Initiative (**HGMI**) (5) from the Genome Sequencing Center at the Washington University in St. Louis. This initiative aims to provide simply annotated, deep draft genome sequences for 100 cultured representatives of the phylogenetic diversity documented by 16S rRNA surveys of the human gut microbiota. From these, 45 projects are already in progress and available in GOLD (the list is available through the search page with the term "Human gut microbiome" as the Relevance search field). The second has been launched earlier this year by the Department of Energy (DOE) - Joint Genome Institute (JGI) and is called Genomic Encyclopedia of Bacteria and Archaea (**GEBA**) (6). GEBA aims the systematic filling in the sequencing gaps along the bacterial and archaeal branches of the tree of life and represents the first systematic attempt to use the tree of life itself as a guide for sequencing target selection. To test the feasibility of a large scale project, DOE-JGI has initiated a pilot project to sequence 100 bacterial and archaeal genomes based on the phylogenetic positions of organisms in the tree of life. The GEBA pilot project is in collaboration with the German Resource Centre for Biological Material (DSMZ) (7) which provides the DNA for the selected organisms. Currently, 79 GEBA projects are reported on GOLD (the list is available through the search page with the term "GEBA" as the Relevance search field).

In addition to the above two large scale sequencing initiatives, a number of National and International efforts for systematic exploration of the Biodiversity have been initiated the last few years, which is also expected to lead to significant increase of sequencing projects. Such efforts include the MikroBioKosmos initiative in Greece (8), the Australian Genome Alliance (9), the Biodiversity Research Initiative in Germany (10), the National BioResource project in Japan (11), the International Census for Marine Microbes (12) and others.

Next to the genome projects, metagenomes and metadata (both for the tracking projects and for the organisms/environments) are the new and fast evolving data types in GOLD and will be discussed in more detail below.

CURRENT STATUS OF THE DATABASE

Published Complete Genomes

GOLD is currently reporting 639 completed genome projects, which is more than double the number since the previous report (3). These are the projects that have their complete sequence deposited to public databases such as GenBank (13), EMBL (14), or DDBJ (15). However, a genome publication is not always available in the literature for these projects since quite often submitters choose to release their sequence data to the community prior of preparing or submitting a publication. This has undoubtedly significantly increased the speed of releasing complete genomes and the entire community benefits from the accelerated availability of the sequences in the public databases. From the 639 complete and published genome projects, 527 are bacterial, 47 are archaeal and 65 are eukaryotic. In the case of several large eukaryotic genomes, the sequencing completion level cannot be the same with that of the microbes, so their sequence status is reported

as Quality Draft (information available in the download file). These are 56 of the 65 eukaryotic projects reported as complete.

Ongoing Genome Projects

In addition to the complete projects, there are currently 2158 ongoing sequencing projects. 1328 of those are bacterial, 59 archaeal and 771 are eukaryotic projects. The latter include 271 EST projects, 74 projects that aim for specific genomic regions or constitute general genome surveys, and 426 whole genome sequencing projects. These can be retrieved by using GOLD's search engine, selecting "EST" or "Genome-Regions" or "Genome-Survey" at the **Type** field.

From the 2158 ongoing projects, 125 are also considered complete at this point, that is the sequencing phase has been completed but the data are not yet submitted to the public sequencing repositories and 513 have already a draft version available. These can be retrieved using the search engine through the **Status** field.

A number of the reported projects (either complete or ongoing) are proprietary and their data may never be released. There are currently 86 such projects reported which can be retrieved by selecting "Proprietary" at the **Availability** field of the Search page. Usually only the information for the sequencing project itself has been made available in these cases.

Metagenome Projects

During the last two years we have witnessed a constantly growing number of metagenomic projects being initiated, and the expectation is that their number will keep on growing as the sequencing technology improves. GOLD is now reporting 108 distinct metagenome projects, 25 of which are considered under a certain criterion complete. For GOLD, the project completion criterion for metagenomes is that the data are deposited in the public databases and the paper describing the project is also published. The organization, structure and presentation of the metagenome data is described in more detail below.

MetaData

Two types of metadata are provided by GOLD: (i) project metadata and (ii) organism/environment metadata. The current status of the different fields and the number of projects with associated data for each of the corresponding fields, is shown on Table 1. Evidently, some of the metadata fields are populated with information for all or most of the projects, while other fields (particularly newer ones such as the pH), are yet to be curated for the majority of the projects.

NEW DEVELOPMENTS

Organization of Metagenomic projects

The project semantics, organization of the data and the presentation of metagenome projects, is still at a very early stage. Given the inherent differences they have compared to the isolate genome projects in most cases there is a need for development of new storing, organization and presentation methods. Some of the main challenges here include: (a) definition of the metagenome project, (b) standardized description of the project name, (c) classification of the metagenome projects, (d) capturing and displaying occasionally large number of distinct samples per project, (e) capturing and displaying number and phylogenetic distribution of the organisms in every sample, (f) capturing and displaying the metadata for individual samples as well as for the entire project, (g) create proper and standardized metadata and capturing them for every sample/project. GOLD will be gradually addressing each of these problems over the next several releases. While the recommendations of the MIGS/MIMS consortium (16) will be in principle adopted for all of the above issues, when ever there is urgency for immediate solutions, there will be novel implementations.

To this extend and in the absence of currently available solutions, the current release of GOLD is mainly addressing the first three problems described above:

(a) Definition of a metagenome project: there has been already a lot of confusion on this, and quite often in the same database for some cases, every sample constitute separate project, while for others, all the samples are grouped under a single project. To avoid such discrepancies, and to group the samples of the same study, a mategenome project in GOLD is considered a single study. All related samples will be presented as individual samples of the same project. For example the project Gm00100 (17), has 13 samples, while Gm00071 has 5 samples.

(b) Standardized description of the project name: this is a already major problem in the field, as quite often the same study (project) is named differently across several different databases. As the number of projects will grow, and several studies with similar focus will appear, it will become very difficult to track the same project across different databases, without a standardized naming convention. An initial effort is made to this direction with the current release, which will be further developed and evolve through the community's feedback. The structure implemented for the metagenome project naming is similar to the Genus-species-strain structure of the isolate genomes and is available from the GOLD CARD pages of each project. Accordingly, each metagenome project name is comprised from up to three types of information: (i) **Project Object** (equivalent to Genus level), which is

describing the habitat (i.e. object) of the community, e.g. Air, Gut, Endophytic, Soil, Wastewater, Hot Spring, Fossil, Marine, etc. (ii) **Project Subject** (equivalent to species), which is describing the location (i.e. subject) of the community, e.g. Human, New York, Neanderthal, etc. and (iii) **Project Identity** (equivalent to strain), which will be describing the specific type (i.e. identity) of the community, e.g. lean and obese, adults, Archaea, etc. This type of naming convention (or others similar to this) will allow avoiding cases where one project would be named New York Air, and another Air from New York or air from Texas. The above structure will not only help grouping based on object, but also on subject. Rather than having all projects grouped under the first word which is the object (e.g. Gut) grouping and retrieval will be also possible based on the subject, (e.g. Human or healthy Human) which will the list all microbiomes based on subject.

(c) Classification of the projects: similar to the two problems described above, a classification schema analogous to the Taxonomic classification available for the isolate organisms, does not yet exists for metagenomes. Again, similar to the approach above, rather than waiting to develop the ultimate classification schema where all possible information or environments could be integrated, we have implemented one, restricted to the projects that are currently available. As new projects will appear that do not fit to the current classification, this will gradually evolve to include the new data. In parallel, when such a schema will be available from the MIMS consortium (16), GOLD will adopt it accordingly. The current metagenome classification is presented in the Information field in the Metagenome table list. All projects are organized under three main categories: (i) **Environmental** (e.g. Environmental-Air, Environmental-Marine, etc.), (ii) **Endosymbiotic** (e.g. Endosymbiotic-Human, Endosymbiotic-Plants, etc.), and (iii) **Synthetic** (e.g. Synthetic-Simulated, Synthetic-Bioreactor, etc.). The GOLD classification for Metagenomes is also available through the Search page, under Phylogeny. This will soon be separated from Organism Phylogeny, to form a distinct Search field only for the Metagenome Classification data.

New Data Fields

In addition of initiating metagenome project tracking and classification schemas, since the last report (3), a number of additional data fields have been added to the database, both in the project tables, as well as in the search engine. These include the fields **(a) Country**, which displays the name of the countries that have genome project. All the projects are currently distributed across 31 countries (including a few multinational efforts); **(b) Sequencing method**, is added to denote if 454 or other methods are used for sequencing; **(c) Sequencing depth** is added when the information is provided; **(d) pH**; **(e) Temperature**, **(f) Project Status** is added to distinguish the completion of sequencing versus the completion of the project; **(g) Metagenome Samples** as described above are also added as a separate field for each of the metagenome projects. In the future these will be further developed to allow the capturing of individual metadata for each of the samples in addition to the metadata for the entire project.

New Pages

A number of new pages have been added. These include: **(a) GOLD CARD** pages for every project, which is available from the link of every GOLD_STAMP ID. The information in every one of these pages is organized into three tables: (i) Organism information, (ii) Genome project information, and (iii) External links. Future developments here will include expanding the information and reorganizing the structure of the three tables closer to the structure shown on Table 1; **(b) Taxonomic Tree** of the projects. Here, the NCBI taxonomy is used to display the number of GOLD sequencing projects down to the Genus level. This is quite helpful in identifying taxonomic groups that are not yet covered from sequencing projects.

Data Availability and interconnectivity

All Data from GOLD are available according to the Creative Commons License of Attribution-NonCommercial-ShareAlike (18). Most of the data can now be downloaded to an excel file, which facilitates distribution and wider use. A number of additional data that are not available either in the project tables or in the search page, are now available directly for download. These include **(a) GreenGenes** IDs (19); **(b) StrainInfo** IDs (20); **(c) GCAT** IDs (21) and **(d) IMG** IDs (22). Accordingly, this file is also providing a mapping across the above resources and those from NCBI (Entrez Project and Taxonomy IDs). Additional fields in this file include the NCBI Taxomomic levels of Superkingdom, Phylum, Class, Order, Family, Genus and Species.

Other data available for download include a regularly updated statistical data file, which is accessible from the Statistics link of the front page (see below).

OVERVIEW STATISTICS

Although several different types of statistics, related to each of the data fields, can be derived from the user at any point using the search engine, or the available for download data, the database also provides readily available graphical

overviews for specific data types. These are provided through the link "**Gold Statistics**" available on the home page of the database, and include the following data types

Sequencing centers

More than half of the 2900 currently available sequencing projects on GOLD are distributed among only four major sequencing centers (since TIGR and the Venter Institute have recently merged). When only the Archaeal and Bacterial projects are taken into account, two sequencing centers alone seem to carry more than half of the world's production. These are the Joint Genome Institute (JGI) and the Venter Institute (JCVI) with TIGR. On top of the list in both cases is the JGI which is the Department of Energy (DOE) sequencing facility with 23% and 27% of world's production respectively (Figure 1). This is based on the number of unique individual projects, and do not correspond in any way with the actual size of the project or the number of sequenced bases which is harder to monitor.

Phylogenetic distribution

The sampling bias towards only three major bacterial lineages (Proteobacteria, Firmicutes and Actinobacteria) continues to persist despite the large increase in sequencing projects as was previously reported (3). As shown on Figure 1, even though the number of Bacterial genome sequencing projects has increased 2.3 fold over the last 2.5 years, the percentage of the three major lineages remains almost entirely unchanged. The development of novel methods that bypass the major restriction of culturing the organism for sequencing (23,24) will hopefully alleviate this bias.

DATABASE AVAILABILITY

GOLD can be accessed at http://www.genomesonline.org/
Further comments and feedback are welcome at mail@genomesonline.org.

ACKNOWLEDGEMENTS

GOLD has been maintained and developed mostly based on the volunteer work of its small team. We are grateful to all the colleagues who kindly provide information for the more accurate monitoring of the genome projects. The support of Tatiana Drakakis and Rashida Lathan, and the continuous contributions of Philip Hugenholtz, Tomer Altman, Krishna Palaniappan and Victor Markowitz, are especially acknowledged. The list of all contributors is available at:
http://www.genomesonline.org/acknowledgments.html
The work presented in this paper was partially supported by the Director, Office of Science, Office of Biological and Environmental Research, Life Sciences Division, U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

REFERENCES

1. Kyrpides, N (1999) Genomes OnLine Database (GOLD 1.0): a monitor of complete and ongoing genome projects world-wide. *Bioinformatics*. **15**(9):773-4.
2. Bernal, A., Ear, U. and Kyrpides, N. (2001) Genomes Online Database (GOLD): A Monitor pf genome projects world-wide. *Nucleic Acid Research* **29**, 126-127.
3. Liolios, K., Tavernarakis, N., Hugenholtz, P. and Kyrpides, N.C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of Genome Projects world-wide. *Nucleic Acids Res* **34**, D332-4
4. Diggle MA, Clarke SC. (2004) Pyrosequencing: sequence typing at the speed of light. *Mol Biotechnol*. **28**(2):129-37
5. http://genome.wustl.edu/hgm/HGM_frontpage.cgi
6. http://mars.jgi-psf.org/programs/GEBA/index.html
7. http://www.dsmz.de/
8. http://www.mikrobiokosmos.org/
9. http://www.genomealliance.org.au/
10. http://www.dfg.de/en/news/press_releases/2006/press_release_2006_25.html
11. http://www.nbrp.jp/index.jsp
12. http://www.coml.org/descrip/icomm.htm
13. Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, and David L. Wheeler.(2007) GenBank *Nucleic Acids Res* 2007 **35**: D21-D25
14. Tamara Kulikova, Ruth Akhtar, Philippe Aldebert, Nicola Althorpe, Mikael Andersson, Alastair Baldwin, Kirsty Bates, Sumit Bhattacharyya, Lawrence Bower, Paul Browne, et al. (2007) EMBL Nucleotide Sequence Database in 2006 *Nucleic Acids Res* 2007 **35**: D16-D20

15. Okubo, K., Sugawara, H., Gojobori, T., Tateno, Y. (2006) DDBJ in preparation for overview of research activities behind data submissions *Nucleic Acids Res*. **34**, D6–D9
16. Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, MJ.,. Angiuolii, SV., et al. (2007) Towards a richer description of our complete collection of genomes and metagenomes: the "Minimum Information about a Genome Sequence" (MIGS) specification. *Nature Biotechnology*, in press.
17. http://genomesonline.org/GOLD_CARDS/Gm00100.html
18. http://creativecommons.org/licenses/by-nc-sa/2.5/
19. DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. (2006). Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. *Appl Environ Microbiol* **72**, 5069-72.
20. Dawyndt P, Dedeurwaerdere T, Swings J. (2007) Exploring and exploiting microbiological commons: contributions of bioinformatics and intellectual property rights in sharing biological information. *International Social Science Journal*, **188**, 249-258
21. http://darwin.nox.ac.uk/gsc/gcat/xtr/genome-catalogue
22. Markowitz VM, Korzeniewski F, Palaniappan K, Szeto E, Werner G, Padki A, Zhao X, Dubchak I, Hugenholtz P, Anderson I, Mavromatis K, Ivanova N, Kyrpides NC. 2006. The Integrated Microbial Genomes (IMG) system. *Nucleic Acids Research*, **34**, D344-D348
23. Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, Szeto E, Platt D, Hugenholtz P, Relman DA, et al. (2007) Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. *Proc Natl Acad Sci U S A*. **104**, 11889-94
24. Podar, M., Abulencia, CB., Walcher, M., Hutchison, D., Zengler, K., Garcia, JA., Holland,T., Cotton, D., Hauser, L. and Keller, M (2007) Targeted Access to the Genomes of Low-Abundance Organisms in Complex Microbial Communities. *Appl Environ Microbiol.* **73**, 3205–3214

Table 1. Metadata types available from GOLD

Project Metadata fields	**No. of projects**	**Organism/Environment metadata**	**No. of projects**
1. GOLD Project ID	2905	1. Domain	2905
2. GCAT ID	2905	2. Phylum	2905
3. NCBI Project ID	1903	3. Class	2905
4. IMG OID	829	4. Order	2905
5. Sequencing Method	797	5. Family	2905
6. Sequencing Coverage	401	6. Genus	2905
7. Project Type	2905	7. Species	2905
8. Sequencing Status	2905	8. Strain	2113
9. Project Status	1375	9. Serovar	177
10. Country	2905	10. Taxon ID	2806
11. Availability	2905	11. StrainInfo ID	320
12. Sequencing center	2896	12. Greengenes ID	707
13. Project Relevance	2241	13. Culture Collection ID	595
14. Funding Center	2108	14. Size	1717
15. Sequence Data	1160	15. Gene Number	991
16. Database	1983	16. Chromosome Number	793
17. Publication	448	17. Plasmid Number	777
18. Release Date	664	18. GC%	1184
19. Contact Name	2158	19. Phenotype	2123
20. Contact Email	2150	20. Habitat	1962
		21. Disease	983
		22. Temperature	626
		23. pH	69
		24. Isolation	1023
		25. Symbiont	122

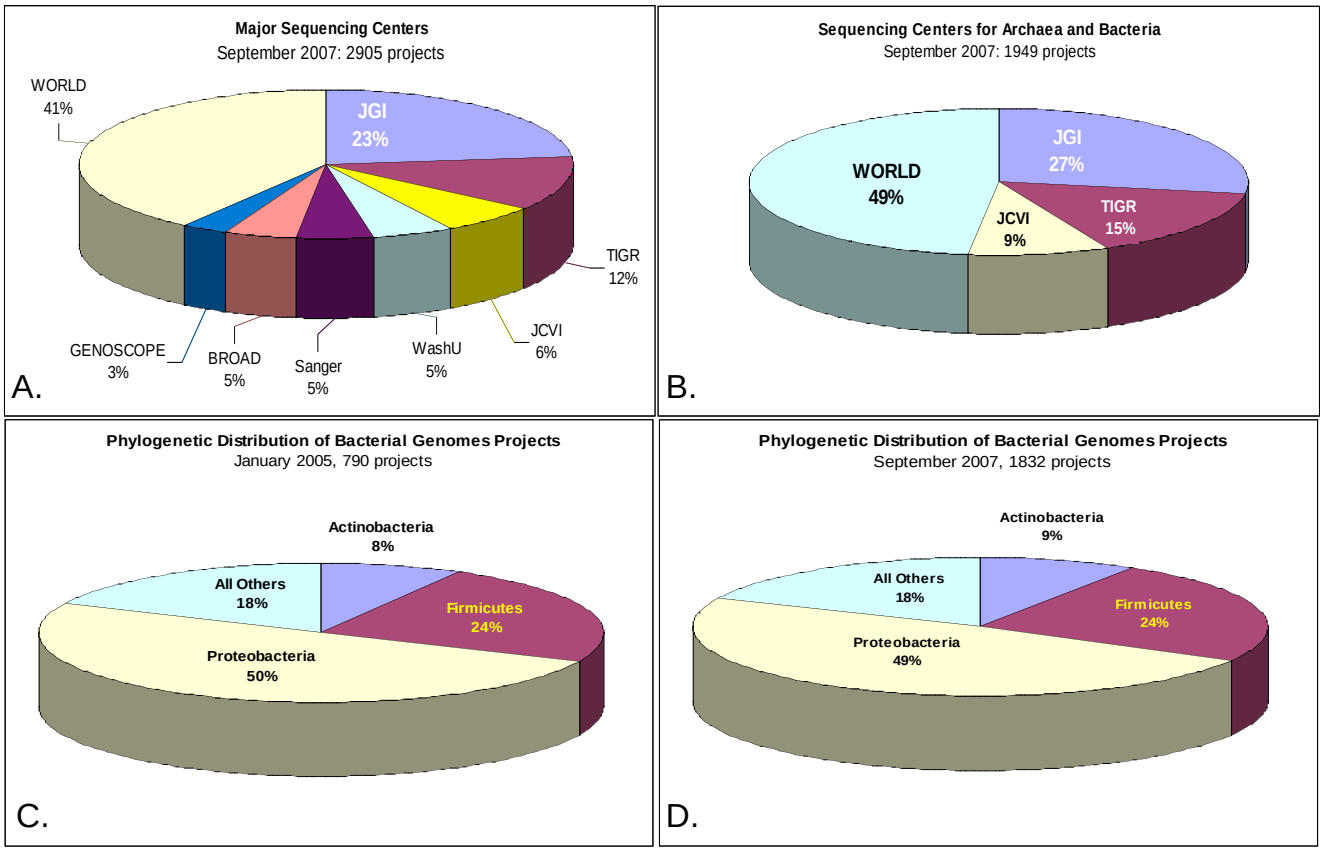

Figure 1. Statistical information available in GOLD. **A.** Distribution of the 2995 genome projects across the major sequencing centers. Abbreviations are for, JGI: Joint Genome Institute, TIGR: The Institute for Genome Research, JCVI: J. Craig Venter Institute, WashU: Washington University. **B.** Distribution of the 1949 Bacterial and Archaeal genome projects across the major sequencing centers. **C.** Phylogenetic distribution of the 790 bacterial genome projects on January of 2005. **D.** Phylogenetic distribution of the 1832 bacterial genome projects on September of 2007.